图书情报工作
第65卷 第23期 2021年12月

科学数据与学术文献关联服务的研究与实现*

黄永文[1] 孙坦[2,3] 赵瑞雪[1,3] 鲜国建[1,3] 李娇[1] 罗婷婷[1]
[1] 中国农业科学院农业信息研究所 北京 100081 [2] 中国农业科学院 北京 100081
[3] 农业农村部农业大数据重点实验室 北京 100081

摘 要：[目的/意义]针对科研人员日益强烈的科学数据检索与发现需求，丰富和完善科学数据的元数据，实现科学数据与学术文献的深度关联发现。[方法/过程]通过对国内外关联服务方式和服务实践进行分析和总结，提出科学数据检索与关联服务系统架构，并实现学术资源元数据采集及融合、科学数据元数据丰富与增强以及科学数据检索与关联发现服务。[结果/结论]科学数据元数据质量的改善可以支持科学数据和学术文献之间更深层次、更细粒度的语义关联服务，助力用户发现科学数据以及与其相关联的学术文献。

关键词：科学数据 学术文献 数据检索 关联发现
分类号：G253
DOI：10.13266/j.issn.0252-3116.2021.23.013

1 引言

随着数据密集型科研范式以及数据科学的兴起与发展，科学数据在科学研究、科技创新、循证决策中的支撑和保障作用愈发明显。在国家层面，科学数据和学术文献资源已被欧美等发达国家归类为国家基础设施的重要组成部分，欧盟、美国、德国等制定了相关战略规划和政策，促进科学数据的共享和可重用。众多知名出版社、基金资助机构、科研机构、学会联盟等纷纷制定科学数据共享政策，出版社明确要求或建议作者在提交论文的同时提交相关的支撑数据，并为文献和数据分别分配永久性唯一标识符（如DOI），同时专门出版科学数据描述论文的数据期刊也应运而生。一些国际重要组织相继推出行动计划和标准框架，如欧洲开放科学云（European Open Science Cloud，ESOC）[1]、文献与数据互连框架（Scholarly Link eXchange，Scholix）[2]、FAIR数据原则[3]、Datacite元数据框架[4]、Elixir互操作规范[5]等，致力于构建开放的、可扩展的、可靠的科学数据基础设施和数据共享生态环境，使科研人员可以轻松地访问和使用科学数据，并呼吁在出版社和数据仓储库之间创建互联机制，促进学术文献、科学数据等资源的访问及关联发现。

与此同时，研究人员也逐渐意识到科学数据、科技文献的某些联系对于提升科研效率的重要性，Elsevier在2019年发布的Trust in Research报告中指出[6]，约57%的研究人员会进一步检查文献的附录数据。将学术文献和科学数据联系起来，可以促进学术文献和科学数据的可发现和可检索，提高科研成果的透明度和可重用性。目前，国际上一些知名的出版商、搜索引擎、数据中心等纷纷推出学术文献和科学数据的关联服务，如PubMed[7]、Elsevier[8]、Web of Science[9]、Scopus[10]、Dimensions[11]、ELIXIR数据平台[12]、TAIR拟南芥信息资源服务平台[13]、科学数据银行（ScienceDB）等都提供了将数据集链接到出版物的服务；谷歌学术[14]、OpenAIRE[15]、RD-Switchboard[16]等建立大型科研图谱，将文献、数据集、作者、机构、项目、资助者等学术实体/科学交流实体联系起来，从而建立全面、连通的数据生态系统。

虽然现有服务系统在一定程度上解决了科学数据的可发现问题，但由于科学数据的分散异构以及元数据质量等问题，科学数据与学术文献的深层次关联还存在不足，并且我国在科学数据元数据质量改善以及

* 本文系国家科技图书文献中心专项"下一代开放知识服务平台关键技术优化集成与系统研发"（项目编号：2021XM55）和中国农业科学院农业信息研究所专项"科技文献与科学数据深度融合及关联发现研究"（项目编号：JBYW-AII-2021-30）研究成果之一。
作者简介：黄永文，副研究馆员，博士；孙坦，副院长，研究员，博士，通讯作者，E-mail：suntan@caas.cn；赵瑞雪，副所长，研究员，博士；鲜国建，研究员，博士；李娇，馆员，博士；罗婷婷，馆员，硕士。
收稿日期：2021-08-01 修回日期：2021-10-21 本文起止页码：116-125 本文责任编辑：易飞

科学数据和学术文献深度语义关联方面的研究相对比较少，缺少实践和应用层面的探索。因此，笔者聚焦于科学数据与学术文献的关联方式，重点分析国外免费的科学数据与学术文献的关联服务，并设计和实现科学数据的检索与关联服务系统，在语义层次上实现科学数据元数据的改善以及与学术文献关联服务，以期为我国图书情报机构开展科学数据和学术文献之间的关联发现服务提供借鉴。

2 相关研究与实践

2.1 科学数据与学术文献的关联研究

近年来，随着科学数据可重用和共享理念的不断深入，关于科学数据与科技文献的关联方式和关联关系构建的研究逐渐增多。杨宁等[17]将科学数据与科技文献的关联方法分为主动关联和被动关联两大类，又将主动关联分为基于元数据的关联、基于引用的关联以及基于语义的关联。姜恩波等[18]将科学数据与科技文献的关联方法分为基于形式的硬关联和基于内容的软关联。笔者在此基础上将科学数据与科技文献的关联方式研究分为 4 种，主要包括：基于唯一标识符的关联、基于引用的关联、基于元数据的关联、基于语义实体的关联。

2.1.1 基于唯一标识符的科学数据与学术文献关联

科学数据、学术文献等研究成果在存储和发布时，通常被分配数字对象标识符，从而使这些研究成果能够被单独发现和引用。科学数据的唯一标识符包括 DOI 标识、数据访问号(如数据库缩写：数据标识符)、ISLI 标识、Handle 标识、PURL 标识、URN 标识、ARK 标识、CSTR 标识等，其中 DOI 和数据访问号是科学数据与学术文献关联最为常用的唯一标识符。涂勇等[19]、孙文佳等[20]探讨了基于 DOI 实现科学数据与科技文献关联的关联方法和关键技术；朱江等[21]研究了基于国际标准关联标识符 ISLI 标准的科技文献和科学数据的关联；德国国家科学技术图书馆[22]也积极开展基于 DOI 的文献与科学数据之间的关联以及基于 ORCID 关联作者的文献和科学数据的探索。

2.1.2 基于引文的科学数据与学术文献关联

科研活动每年都会产生大量类型多样的科学数据，这些科学数据被不同团体和科研人员使用，并在出版物中加以引用，由此产生了科技文献与其支撑数据之间的关联。这种关联关系不但使得科学数据具备了重用性，也使得科技文献与科学数据产生了关联关系，一些学者对基于引文的科学数据与学术文献关联方式和方法进行研究，并对特定领域文献中的数据集进行识别与抽取。郭学武[23]将基于引文的关联分为直接引用关联、同被引关联以及基于引文的扩展关联 3 种形式；张鑫等[24]以高能物理领域为例，研究基于引文探针的文献与数据的关联算法，通过对关联度的计算发现隐含关联关系。作者经常会在正文中引用科学数据，因此一些学者对从论文全文中识别科学数据进行了探索。N. Riedel 等[25]利用文本挖掘算法，从生物医学领域的文献中检测科学数据引用和可用性语句；L. L. Hou 等[26]提出数据集实体识别模型 MDER，并从论文全文内容中提取引用和提及的数据集，并在计算机科学领域进行了验证；B. Ghavimi 等[27]利用半自动的方法识别社会科学领域文献中引用的科学数据集。

2.1.3 基于元数据的科学数据与学术文献关联

基于元数据的科学数据与学术文献关联主要是利用科学数据和学术文献外部及内部特征的相似性而建立的一种关联关系。孙志茹等[28]以生物信息领域的数据描述和文献描述的相似性为出发点进行分析，并提出基于硬连接、基于近邻关系、基于数据聚类和基于主题 4 种数据与文献的关联方式；黄筱瑾认为科学数据和科技文献的元数据关联模式包括作者关联、学科分类号关联、关键词关联[29]，并从数据和文献的元数据描述中提取出表达内容特征的元数据项，利用向量空间模型计算出数据与文献之间的关联关系[30]；贺姝祎等[31]通过分析天文领域科学数据与科技文献在元数据方面的相似性及差异，并基于数据挖掘技术探讨天文领域科学数据与科技文献关联的可行性。

2.1.4 基于语义实体的科学数据与学术文献关联

主要是从语义内容角度实现科学数据与学术文献之间的关联，语义实体是指在科学数据的元数据描述中包含的关键概念、术语或实体(如物种名称、基因名称、蛋白质名称、化学物质、疾病等)，科学数据中语义实体识别的方法主要有作者标注和自动文本挖掘[32]。孙巍[33]运用基于分面分类的描述方法对科学数据实体进行细粒度描述，并在农业小麦育种领域验证科学数据与科技文献实体间的语义关联；丁培[34]提出科学文献和科学数据的细粒度内容语义关联模型，并验证基于本体的实体识别实现数据与文献关联的可行性；T. Clark[35]提出了基于实体和基于论点两种互补的生物医学文献与数据关联集成模型；H. Cousijn 等[36]基于表格内容、生物实体等多种方式建立文献与科学数据的关联关系；I. J. Aalbersberg 等[37]利用文本挖掘、术语提取等技术从全文中挖掘语义实体，并建立语义实

体到数据仓储中数据的关联链接。

2.2 科学数据与学术文献的关联服务实践

国内对科学数据与学术文献之间的关联服务实践刚刚起步，因此笔者主要选取国外免费的科学数据与学术文献的关联服务系统进行分析。DataCite[38] 于 2015 年 8 月推出 DataCite Search 数据搜索工具，为用户提供一站式数据检索服务，并在出版社和数据仓储库之间创建了互联机制；谷歌[39] 认识到数据的重要性日益增加，于 2018 年推出了 Google Dataset Search，将数据集的可发现提升到一个新的水平；研究数据联盟（RDA）、世界数据系统（WDS）[40] 等从数据中心、期刊出版商和研究机构收集数据和文献之间的链接，并推出 ScholeXplorer 数据文献互连服务；美国国立卫生研究院（NIH）[7] 依托《美国国家医学图书馆 2017－2027 战略计划》的行动方案，研发并推出了基于 PMC 的数据关联发现服务；OpenAIRE[41] 通过自动推理建立不同学科的数据集和出版物之间的语义关系，将数据集和出版物进行聚合关联，为科研全过程提供基础及附加服务；Dryad[42]、PANGAEA[43]、HEPData[44] 等领域科学数据存储库也已经与文献实现了互联互通，在科学数据存储库中检索到的科学数据，除了显示科学数据的基本信息外，还提供了与文献的链接。这些免费的科学数据与学术文献的关联服务系统对比情况如表 1 所示：

表 1 免费的科学数据与学术文献关联服务系统对比

名称	类型	年份	国别/地区	学科领域	资源类型	数据集数量/万个	数据接口
DataCite Search	科学数据搜索引擎	2015	德、美、英等 42 个国家/地区	综合	数据集、图片、音视频、软件、模型、工作流、数据论文等	2 641	OAI-PMH API
Google Dataset Search	科学数据搜索引擎	2018	美国	综合	数据集	2 500	无
RDA/WDS ScholeX-plorer	综合搜索引擎	2015	欧盟、美国、澳大利亚等	综合	数据集、文献等	958.7	REST API
NIH PMC	综合搜索引擎	2018	美国	生物、医学	数据集、文献等	--	OAI-PMH API、RESTful API
OpenAIRE EXPLORE	综合搜索引擎	2019	欧盟	综合	数据集、文献、项目、软件等	1187.7	HTTP API
Dryad	领域科学数据存储库	2008	美国	生物	数据集	4.2	OAI-PMH API
PANGAEA	领域科学数据存储库	1993	德国	环境和地球科学、生命科学	数据集、项目等	40.4	OAI-PMH API
HEPData	领域科学数据存储库	1975	英国	高能物理	数据集（图片和表格）	9.9	RESTful API

注：①因为在网站上没有找到 PANGAEA 和 HEPData 推出数据与文献关联服务的具体时间，所以表中列出的时间为数据库的建立时间；②国别是指研发或维护该服务系统的机构或联盟所属的国别

综上可以看出，主流的搜索引擎、数据中心等都开始关注科学数据的收集和汇聚，专门针对科学数据进行检索发现，并将文献与科学数据联系起来。数据存储库也积极开展从科学数据到文献的关联链接服务，特别是在生物、物理、医药等自然科学领域，科学数据与学术文献之间的关联和链接服务实践较为成熟，实现了文献与数据之间深层的语义关联服务，而社会科学领域的关联服务实践则较少。越来越多的出版商与数据中心、机构团体开展合作，建立协作共享机制，积极实现数据和文献之间的互通互联，并注重互操作性，如提供数据访问的标准接口（如 OAI-PMH API、Restful API 等）、遵循 FAIR 数据原则等，确保公共存档的数据和文献之间可引用和可关联，有效提高科学数据的可检索性、发现性、可解释性和可重用性。

3 科学数据检索与关联服务系统设计与实现

近年来，科研人员关注的对象不再局限于期刊论文、会议论文、科技报告等文献资源，科学数据也逐渐成为科研人员所需要的重要资源。科研人员会从学术文献出发，从文献的内容中或参考文献中发现科学数据的线索。因此，如何将科学数据和学术文献进行有效的关联，对于提高科研活动效率、加强科学数据的复用和共享、实现更深层次的知识发现，都具有极为重要的现实意义。借鉴国外相关服务系统，笔者围绕如何有效检索科学数据以及如何利用科学数据与学术文献之间关系强化发现服务这一核心问题，挖掘科学数据和学术文献中的语义实体，丰富和完善科学数据元数

据质量，增强科学数据的可发现能力，并将科学数据与学术文献进行深度融合和语义关联，基于唯一标识符、引用关系、元数据和语义实体，将科学数据与相关学术文献联系起来，实现多层次的增强型发现与关联服务，帮助用户快速地发现科技资源。科学数据检索与关联服务系统的架构如图 1 所示：

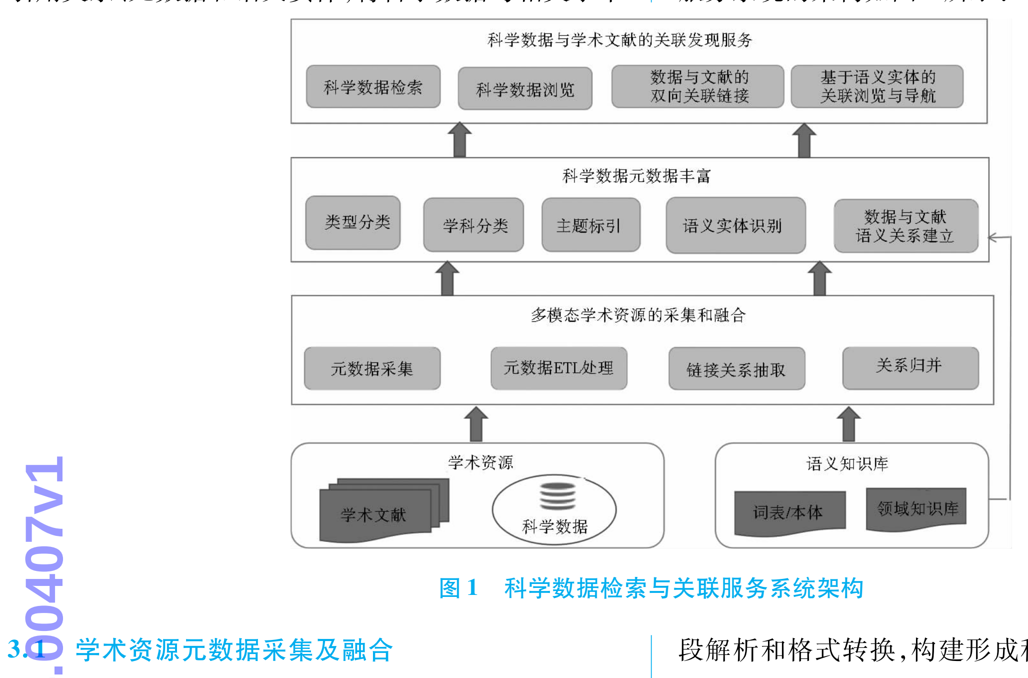


图 1　科学数据检索与关联服务系统架构

3.1　学术资源元数据采集及融合

Datacite 是科学数据唯一标识 DOI 的注册机构，CrossRef 是学术文献唯一标识的注册机构，两者收录了权威的科学数据和学术文献的元数据，因此笔者选取 Datacite 和 CrossRef 作为数据来源。通过 OAI API 方式获取了 Datacite 和 CrossRef 的元数据，利用 Kettle 工具分别进行 ETL 数据处理，并按照一定规则进行数据筛选，只对含有关联关系的科学数据元数据进行字段解析和格式转换，构建形成科学数据元数据库和学术文献元数据库，共包含 376 万余条科学数据元数据，全部带有 DOI 标识符。同时，对科学数据与学术文献、科学数据与科学数据之间的链接关系进行解析、抽取、查重和合并，并基于数据关联模型（见图 2）和 Datacite 元数据框架定义的关系类型受控词汇表（如“IsCited-By”“IsSupplementTo”“HasPart”“IsDerivedFrom”等），形成包含 858 万余条关系记录的关联关系库。

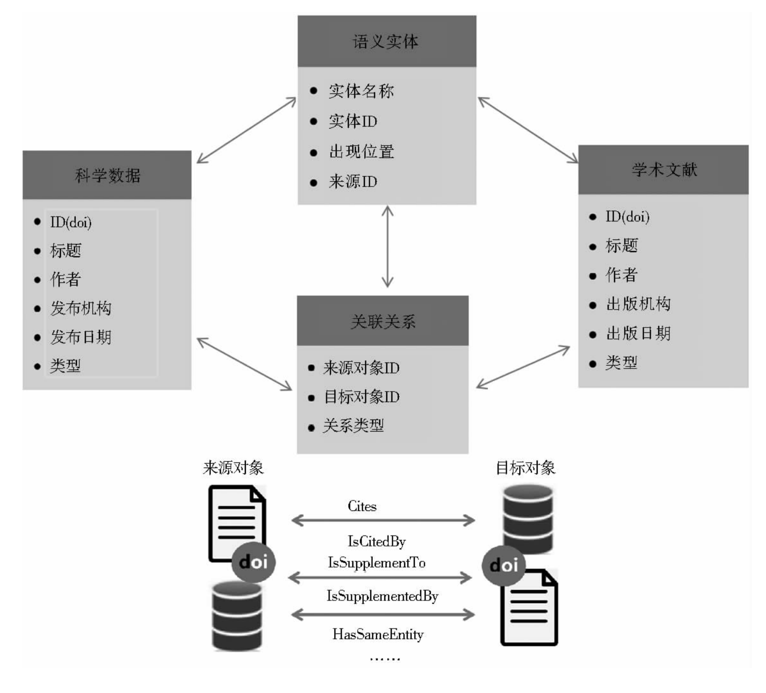


图 2　数据关联模型

在数据处理过程中,采用基于规则的方式对CrossRef参考文献中的科学数据引用记录进行识别和保存,例如参考文献的题名以"Data from:"或"Data for:"开头,共识别和提取了6 483条引用数据,在关联关系库中为相关的科学数据增加"IscitedBy"关系类型。虽然,目前参考文献中出现科学数据引用的数量还比较少,但随着科学数据引用被科研人员广泛认可以及出版社和基金资助机构提出强制要求或建议,更多的研究人员会提交和共享数据,科学数据引用的数量将逐渐增长。截至2020年底,已有13 000多个期刊支持数据提交和共享政策[45],促进科学数据和学术文献链接关系的建立在出版物提交系统以及科学研究生命周期的上游得到实施。同时,随着Crossref和DataCite合作的深入开展,将会进一步确保科学数据和学术文献之间引用关系的完整性和准确性。

本研究形成的关联关系库主要包括科学数据与学术文献之间的关系和科学数据与科学数据之间的关系,其中数量排名前10的关联关系情况见表2,关系类型说明中的A和B至少有一个为科学数据。从表2可以看出,关系类型最多的是"IsPartOf"(占比28.9%)和"HasMetadata"(占比24.4%),"IsPartOf"表示科学数据A是科学数据B的一部分,"HasMetadata"表示科学数据A具有其他元数据B;其次是"IsCitedBy"(占比8.0%)和"IsSupplementTo"(占比6.4%),"IsCitedBy"表示B在引文中包含A,"IsSupplementTo"表示A是B的补充。"IsCitedBy""IsSupplementTo""References""Cites"等是实现科学数据和学术文献关联的主要关系类型。

表2 排名前10的关联关系情况

序号	关系类型	关系数量	所占比例/%	关系类型说明
1	IsPartOf	2 478 868	28.9	表示A是B的一部分
2	HasMetadata	2 094 643	24.4	表示A具有其他元数据B
3	IsCitedBy	689 189	8.0	表示B在引文中包含A
4	IsSupplementTo	550 301	6.4	表示A是B的补充
5	IsDocumentedBy	468 428	5.4	表明B是关于或解释A的文档
6	HasVersion	436 383	5.1	表示A有一个版本B
7	IsVersionOf	429 824	5.0	表示A是B的一个版本
8	IsPreviousVersionOf	367 800	4.3	表示A是B的早期版本
9	References	279 294	3.3	表示B作为A的信息源
10	Cites	118 274	1.4	表示A在引文中包含B

3.2 科学数据元数据丰富与增强

对科学数据进行描述和组织是科学数据共享、检索和利用的前提,元数据可以用于描述科学数据的内容及形式等特征。不过,目前大多数科学数据的元数据信息都比较少,数据描述通常不完整或没有包含足够的详细信息,缺少主题、分类等。笔者在分析科技数据内容特征元数据的基础上,基于词表、本体等知识组织体系,利用科学数据的标题、关键词、摘要(或描述),实现了科学数据的自动分类、主题标引和语义实体标签生成,为科学数据补充主题概念、《中国图书馆分类法》分类号(以下简称"中图分类号")、生物物种实体标签、化学物质实体标签、基因实体标签等信息,增强和丰富了科学数据的语义元数据,为科学数据的分面浏览和关联发现服务提供支撑。科学数据元数据丰富与增强的流程如图3所示:

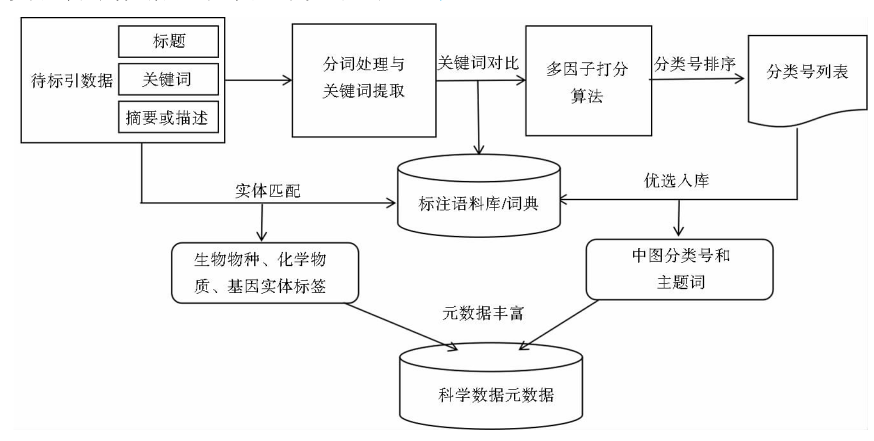


图3 科学数据元数据丰富与增强的流程

3.2.1 语义实体标签的自动生成

语义实体主要指科学数据的标题、关键词、摘要或描述等信息中包含的有意义的实体名称,如物种名称、化学物质、基因名称、蛋白质名称等。笔者主要采用基于词典的方法进行语义实体抽取。首先,在ISTI、Species 2000、Uniport、Mesh等基础上构建领域实体词典,

然后在题名、关键词、描述或文摘中，对词典中已有的实体进行匹配，如果匹配上，则对该实体进行识别和抽取，并标记出现的位置，为该科学数据自动生成相关的实体标签。笔者主要为科学数据增加了化学物质、生物物种和基因实体标签，科学数据的语义实体标签抽取示例如表3所示：

表3 科学数据的语义实体标签抽取示例

序号	题名	DOI	化学物质标签	物种标签	基因标签
1	Supplementary Material for: Involvement of Heat Shock Proteins in Candida albicans Biofilm Formation	10.6084/m9.figshare.5125192.v1	Proteins	Candida	HSP90
2	Data from: A fat-derived metabolite regulates a peptidergic feeding circuit in Drosophila	10.5061/dryad.8hm82	Neuropeptide Y; sepiapterin;	Drosophila; Drosophila melanogaster;	BH4
3	A simple and versatile authenticity assay of coffee products by single-nucleotide polymorphism genotyping	10.6084/m9.figshare.8174558.v1	Coffee; DNA;	Coffea	rbcL
4	Data from: Discrimination of grasshopper (Orthoptera: Acrididae) diet and niche overlap using next-generation sequencing of gut contents	10.5061/dryad.r8c3b	DNA	Chortophaga; Dissosteira; Melanoplus;	rbcL
5	The role of PI3K/Akt signal pathway in the protective effects of propofol on intestinal and lung injury induced by intestinal ischemia/reperfusion	10.6084/m9.figshare.7743152.v1	Malondialdehyde; Propofol; Superoxide Dismutase;	-	PI3K

为了在学术文献和科学数据之间建立基于语义实体的关联关系，除了在科学数据元数据中进行语义实体的识别和抽取，还需要在学术文献的元数据中进行语义实体的识别和抽取。

3.2.2 科学数据的自动分类标引

笔者在文献[46]中结合加权策略提出一种全流程的基于多因子算法的自动分类方法，该方法本身无领域和处理对象限制。基于人工标引经验和训练语料两者优势，结合分类号出现概率、关键词位置权重、命中分类号下各关键词占比、命中分类号下所有关键词出现的频率等多个影响因素，实现自动分类。通过继承复用已有权威语料库（如英文超级科技词表STKOS、中国农业科学叙词表CAT等）、基于高质量权威来源文献数据抽取关键词和学科分类号等多种方式构建标注语料库，即在包含词语、概念、术语等表征文献内容的知识元词库基础上，纳入揭示领域特征的学科分类号，建立主题词-分类号对照数据库，保障后续自动分类的准确性。在采用权威标注语料库和不介入人工审核，仅通过多因子算法计算的情况下，所提出的自动分类方法针对多学科领域学术文献随机样本的准确率和F值均在80%以上。

笔者采用文献[46]中的多因子算法对科学数据进行自动分类。首先，通过分词工具对待标引科学数据的元数据信息（标题、摘要或描述及关键词）进行切词并提取关键词，获取主题信息。然后将提取的关键词与选定的标注语料库中关键词进行完全匹配，获取命中的关键词及相应学科分类号信息，并计算出关键词对应的各学科分类号在该语料库所有学科分类号中的频率。最后，基于学科分类号出现概率与抽取关键词位置权重、命中学科分类号下各关键词在该分类号对应的所有关键词中的占比、命中学科分类号对应所有关键词在该分类号下出现的频率等，进行加权计算并对学科分类号得分排序，选取排名前5的学科分类号作为科学数据的分类号。科学数据自动分类的示例见表4。

3.3 科学数据检索与关联发现服务

基于语义增强的科学数据元数据库，笔者实现了科技数据的检索、浏览和关联发现服务（网址为：http://www.agriknow.cn/nstl/datacite.html），支持按发布日期、类型、生物物种实体、化学物质实体、基因实体、学科分类（中图法）、关键词、发布者、来源（发布机构）、访问许可、资助机构等多角度的浏览，提供科学数据下载链接。并与学术文献检索进行集成，提供“主题词”“支撑数据”等分面的限定检索，实现科学数据与学术文献的双向关联，支持科学数据与学术文献的关联发现和导航服务，并基于语义实体为用户提供更多的研究线索。

对科学数据相关的文献和支持文献的科学数据进行有效集成和关联，用户在检索结果页面上，通过选择左侧分面“支撑数据”，可以浏览关联科学数据的学术文献（见图4）。用户在浏览和查看学术文献详细信息时，通过点击页面右侧的“相关数据”可以直接链接至文献中提及的数据集（见图5右下），也可以在查看科学数据的详细信息时点击“相关文献”，链接至与该科学数据相关的学术文献（见图5左上）。

表 4 科学数据自动分类的示例

序号	题名	DOI	中图分类号	类名
1	Data from: Plant dispersal in the sub-Antarctic inferred from anisotropic genetic structure	10.5061/dryad.4f1r5vg8	Q94;Q;Q14;Q1;S6	植物学;生物科学;生态学(生物生态学);普通生物学;园艺
2	Data from: Root vertical distributions of two Artemisia species and their relationships with soil resources in the Hunshandake desert, China	10.5061/dryad.9zw3r22b4	Q94;S;Q;Q95;X	植物学;农业科学;生物科学;动物学;环境科学、安全科学
3	Data for: Predicting the Impacts of Mutations on Protein-Ligand Affinity Based on Molecular Dynamics Simulations and Machine Learning Methods	10.17632/nwmyvyyy2v.3	Q5;Q7;R3;Q;Q93	生物化学;分子生物学;基础医学;生物科学;微生物学
4	Data for: Seat Allocation Model for High-speed Railway Passenger Transportation Based on Flexible Train Composition	10.17632/466djt2ccv.1	U292. 5; TP311. 52; TV213. 4; TP311. 13; TB383;	铁路通过能力、运送能力;软件开发;高速铁路;数据库理论与系统;特种结构材料
5	Detection and characterization of quantitative trait loci for coleoptile elongation under anaerobic conditions in rice	10.6084/m9.figshare.11993403	Q94;Q;R3;R9;Q7	植物学;生物科学;基础医学;药学;分子生物学

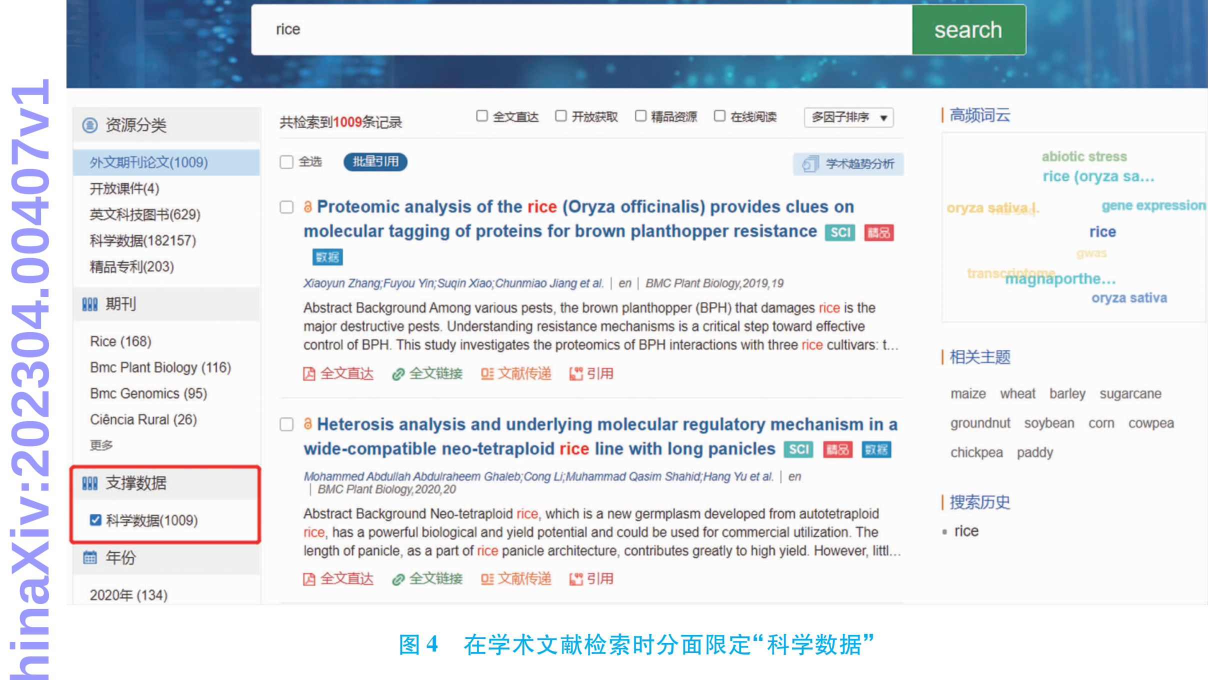


图 4 在学术文献检索时分面限定“科学数据”

笔者从科学数据和学术文献元数据中识别出生物物种名称、化学物质名称、基因名称等,形成语义实体关联标签库,基于这些语义实体标签实现了科学数据与学术文献的深层次关联服务。以生物物种名称、化学物质名称、基因名称等语义实体为起点,可以关联相关的科学数据和学术文献以及链接到 NCBI、Uniprot 等外部数据源。在科学数据的详细信息页面上,分别列出“物物种名称标签”“化学物质名称标签”和“基因标签”,通过点击这些标签可以关联到含有该实体名称标签的科学数据和学术论文,提供列表方式和可视化方式两种展示方式。基于语义实体的关联服务见图 6。

4 结语

近年来,科学数据的数量呈指数增长,科学数据开放、共享、重用的理念已经为研究人员广泛接受。科学数据可以对以论文形式发表的成果进行补充说明,能够帮助用户更加了解科研的整个过程,可用于研究再现及证伪。查找和发现数据是能够重用科学数据的必要前提,笔者基于 Datacite 和 CrossRef 数据构建了科学数据元数据库和关联关系库,设计并实现了科学数据检索以及与学术文献的关联服务系统,基于数据外部特征和语义特征实现科技资源之间深度融合关联,并利用语义实体创建文献与数据之间的相关关系,支持更深层次、更细粒度的语义关联服务,帮助用户从多途径发现科学数据以及与其相关联的学术文献。不过,本研究也存在一些不足之处,如仅仅采用基于词表的方式,对于一些有多个含义的词进行抽取和分类,会导致歧义性或错误的分类。如 Latex 一词,表示乳胶,属于化学物质实体,同时 Latex 也是一种数据格式。因此,未来的研究中将继续优化语义实体的抽取方式,通

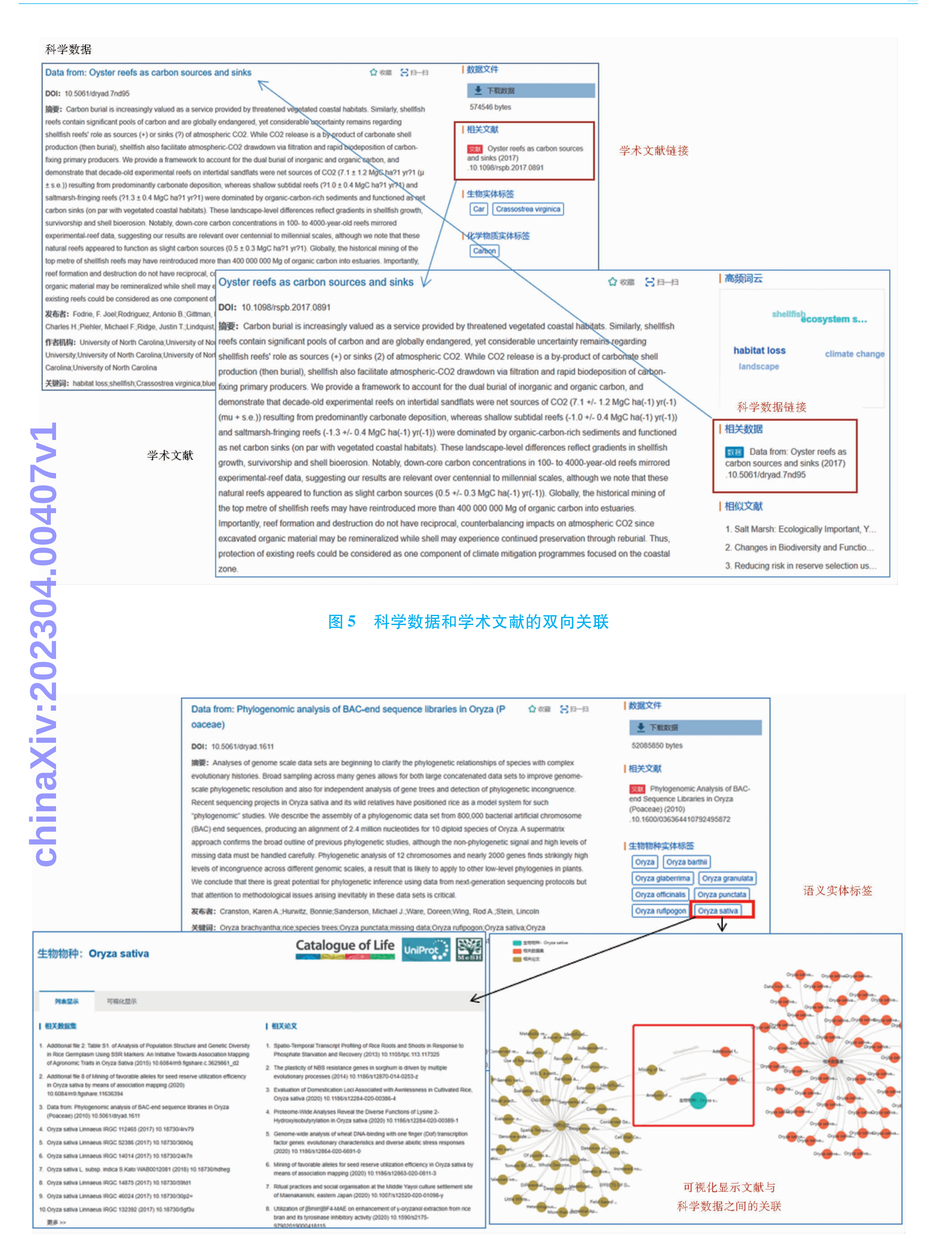


图5 科学数据和学术文献的双向关联

图6 基于语义实体的关联服务

过深度学习的方法,避免该类问题的出现。同时,还将针对关联服务展开更深入的研究,如对文献中的数据访问控制号进行识别和链接以及对更多类型的语义实体进行识别,如蛋白质名称、病虫害名称等,对科学数据和学术文献之间关系进行计算和推理,发现数据和文献之间的更深层次的关联关系。

参考文献:

[1] European Open Science Cloud[EB/OL]. [2020-10-20]. https://eosc-portal.eu/about/eosc.
[2] BURTON A, KOERS H, MANGHI P, et al. The Scholix framework for interoperability in data-literature information exchange[J/OL]. [2020-10-20]. http://www.dlib.org/dlib/january17/burton/01burton.html.
[3] WILKINSON M D, DUMONTIER M, AALBERSBERG I J J, et al. The FAIR guiding principles for scientific data management and stewardship[J]. Scientific data, 2016:160018. https://doi.org/10.1038/sdata.2016.18.
[4] DataCite metadata schema[EB/OL]. [2020-10-20]. https://schema.datacite.org/.
[5] Elixir. Interoperability platform[EB/OL]. [2020-10-20]. https://elixir-europe.org/platforms/interoperability.
[6] Elsevier. Trust in research[EB/OL]. [2020-10-20]. https://www.elsevier.com/connect/trust-in-research.
[7] Discovering associated data in PMC[EB/OL]. [2020-10-20]. https://ncbiinsights.ncbi.nlm.nih.gov/2018/11/15/discovering-associated-data-in-pmc/.
[8] Elsevier. Linking research data and research articles on ScienceDirect[EB/OL]. [2020-10-20]. https://www.elsevier.com/authors/tools-and-resources/research-data/data-base-linking.
[9] Web of Science. Data Citation Index[EB/OL]. [2020-10-20]. https://clarivate.com/webofsciencegroup/solutions/webofscience-data-citation-index/.
[10] Scopus. Data linking[EB/OL]. [2021-03-10]. https://blog.scopus.com/topics/data-linking.
[11] Dimensions. Linked research data from idea to impact[EB/OL]. [2021-03-10]. https://www.dimensions.ai/.
[12] Elixir data platform[EB/OL]. [2021-03-10]. https://elixir-europe.org/platforms/data.
[13] GARCIA-HERNANDEZ M, BERARDINI T Z, CHEN G H, et al. TAIR: a resource for integrated Arabidopsis data. Functional & integrative genomics, 2002,2(6):239-253.
[14] SULLIVAN D. A reintroduction to our knowledge graph and knowledge panels[EB/OL]. [2020-10-20]. https://www.blog.google/products/search/about-knowledge-graph-and-knowledge-panels/.
[15] OpenAIRE. OpenAIRE - research graph[EB/OL]. [2020-10-20]. https://graph.openaire.eu.
[16] RD-Switchboard[EB/OL]. [2020-10-20]. https://www.rd-switchboard.org.
[17] 杨宁,文奕,张鑫,等. 高能物理科学数据与科技文献关联研究[J]. 图书馆学研究,2019(1):47-52.
[18] 姜恩波, 裴玉香. 科学文献与科学数据的融合方法与实例研究[J]. 知识管理论坛,2019,4(2):69-79.
[19] 涂勇,彭洁. 基于DOI技术的科学数据与科技文献融合的研究[J]. 数字图书馆论坛,2007(10):28-31.
[20] 孙文佳,常娥. 科学数据与科技文献关联分析[J]. 图书馆理论与实践,2017(3):49-53.
[21] 朱江,李欣怡,姜恩波,等. 基于ISLI标准的科技文献和科学数据的关联[J]. 图书馆理论与实践,2020(5):80-83,91.
[22] KRAFT A, DREYER B, LOWE P, et al. 14 Years of PID services at the German National Library of Science and Technology (TIB): connected frameworks, research data and lessons learned from a national research library perspective[J]. Data science journal, 2017,16(36):1-10.
[23] 郭学武. 基于引文的科学数据与科技文献关联研究[J]. 情报科学,2014,32(4):59-62.
[24] 张鑫,文奕,杨宁,等. 基于引文探针的文献与数据的关联算法与应用——以高能物理领域为例[J]. 情报理论与实践,2019,42(10):151-156.
[25] RIEDEL N, KIP M, BOBROV E. ODDPub- a text-mining algorithm to detect data sharing in biomedical publications[J]. Data science journal,2020,19(42):1-14.
[26] HOU L L, ZHANG J, WU O, et al. Method and dataset entity mining in scientific literature: a CNN + Bi-LSTM model with self-attention[EB/OL]. [2021-10-08]. https://arxiv.org/abs/2010.13583.
[27] GHAVIMI B, MAYR P, LANGE C. et al. A semi-automatic approach for detecting dataset references in social science texts[J]. Information services & use,2016, 36(3/4): 171-187.
[28] 孙志茹, 韩涛, 杨文. 生物信息学科学数据与科学文献的关联关系分析[J]. 图书情报工作, 2008,52(2):88-91.
[29] 黄筱瑾. 基于元数据的科学数据与科技文献关联研究[J]. 情报理论与实践,2013,36(7):27-30.
[30] 黄筱瑾. 基于内容特征的科学数据与科技文献关联研究[J]. 现代情报,2018,38(1):56-59.
[31] 贺姝祎,魏韧,吴茂春,等. 科技文献与观测数据的关联性在天文领域的应用研究[EB/OL]. [2021-09-10]. https://d.wanfangdata.com.cn/conference/8469846.
[32] 卫军朝. 科学文献与科学数据关联实践研究——以Elsevier为例[J]. 国家图书馆学刊,2017,26(3):93-101.
[33] 孙巍. 科学数据与科技文献关联发现系统研究与实现[EB/OL]. [2021-09-10]. https://d.wanfangdata.com.cn/conference/7611510.
[34] 丁培. 科学文献与科学数据细粒度语义关联研究[J]. 图书馆论坛,2016,36(7):24-33.
[35] CLARK T. Argument graphs: literature-data integration for robust and reproducible science[EB/OL]. [2021-01-20]. http://

www. isi. edu/ikcap/sciknow2015/papers/Clark. pdf.

[36] COUSIJIN H, HAAK W, KOERS H. Finding better ways to connect research data with scientific literature[EB/OL]. [2021 – 01 – 20]. https://www. elsevier. com/connect/finding-better-ways-to-connect-research-data-with-scientific-literature.

[37] AALBERSBERG I J, KAHLER O. Supporting Science through the Interoperability of Data and Articles[EB/OL]. [2021 – 10 – 08]. http://www. dlib. org/dlib/january11/aalbersberg/01aalbersberg. html.

[38] DataCite Search[EB/OL]. [2021 – 03 – 10]. https://search. datacite. org/.

[39] Google Dataset Search[EB/OL]. [2021 – 03 – 10]. https://datasetsearch. research. google. com/.

[40] ScholeXplore[EB/OL]. [2021 – 03 – 10]. https://scholexplorer. openaire. eu/.

[41] OpenAIRE explore[EB/OL]. [2021 – 03 – 10]. https://explore. openaire. eu/.

[42] DRYAD[EB/OL]. [2021 – 03 – 10]. Our platform. https://datadryad. org/stash/our_platform.

[43] Elsevier and PANGAEA Link Contents for easier access to full earth system research[EB/OL]. [2021 – 03 – 10]. https://www. elsevier. com/about/press-releases/science-and-technology/elsevier-and-pangaea-link-contents-for-easier-access-to-full-earth-system-research.

[44] HEPData[EB/OL]. [2021 – 03 – 10]. https://www. hepdata. net/.

[45] STM. Research data share-link-cite[EB/OL]. [2020 – 10 – 20]. https://www. stm-researchdata. org/.

[46] 李娇,黄永文,罗婷婷,等. 基于多因子算法的自动分类研究[J]. 数据分析与知识发现,2020,4(11):43 – 51.

作者贡献说明:

黄永文:撰写与修改论文;
孙坦:提出论文写作思路,论文定稿;
赵瑞雪:论文最终版本修订;
鲜国建:采集及处理数据;
李娇:收集资料,调研与分析文献;
罗婷婷:收集资料,调研与分析文献。

Research and Implementation of Linking Services Between Scientific Data and Academic Literature

Huang Yongwen[1] Sun Tan[2,3] Zhao Ruixue[1,3] Xian Guojian[1,3] Li Jiao[1] Luo Tingting[1]

[1] Agricultural Information Institution, Chinese Academy of Agricultural Sciences, Beijing 100081

[2] Chinese Academy of Agricultural Sciences, Beijing 100081

[3] Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081

Abstract: [**Purpose/significance**] To meet researchers' increasing demands for scientific data retrieval and discovery, this study conducts a research on the improvement of the metadata of scientific data and further realizes the in-depth linking discovery between scientific data and academic literature. [**Method/process**] Based on the investigating the methods and experiences of linking service, this study proposed a system architecture of scientific data retrieval and linking services, and realized collection and integration of academic resource metadata, enrichment and enhancement of scientific data metadata, and retrieval and linking discovery services of scientific data. [**Result/conclusion**] The improvement of the quality of scientific data metadata can support deeper and more fine-grained semantic linking services between scientific data and academic literature, and help users discover scientific data and its associated academic literature.

Keywords: scientific data academic literature data retrieval data linking discovery